

Semaine 03

APPRENTISSAGE NON-SUPERVISÉ

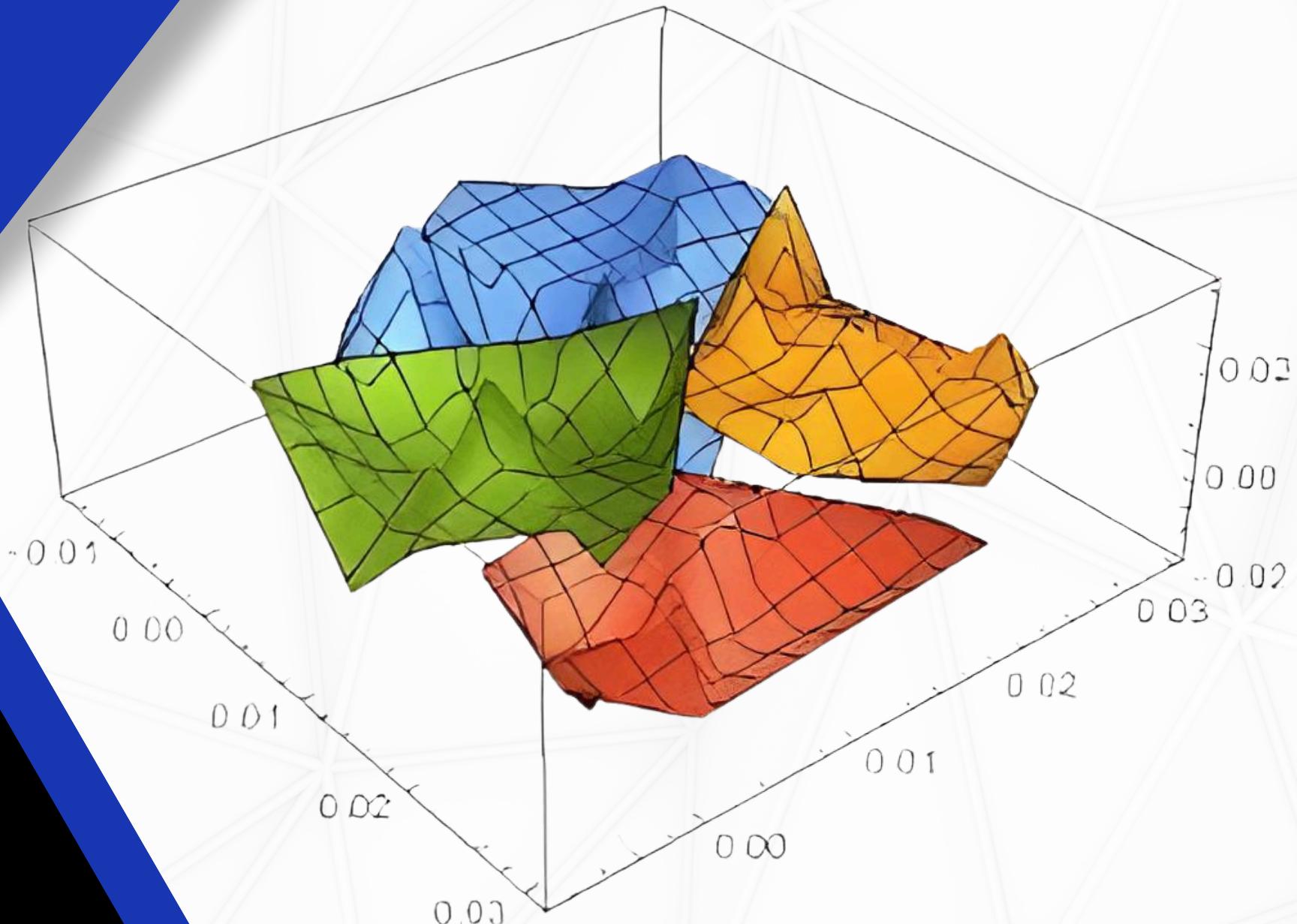


Table Of Content

01. Overview de l'Apprentissage non Supervisé

02. Algorithmes Non supervisé

03. Clustering

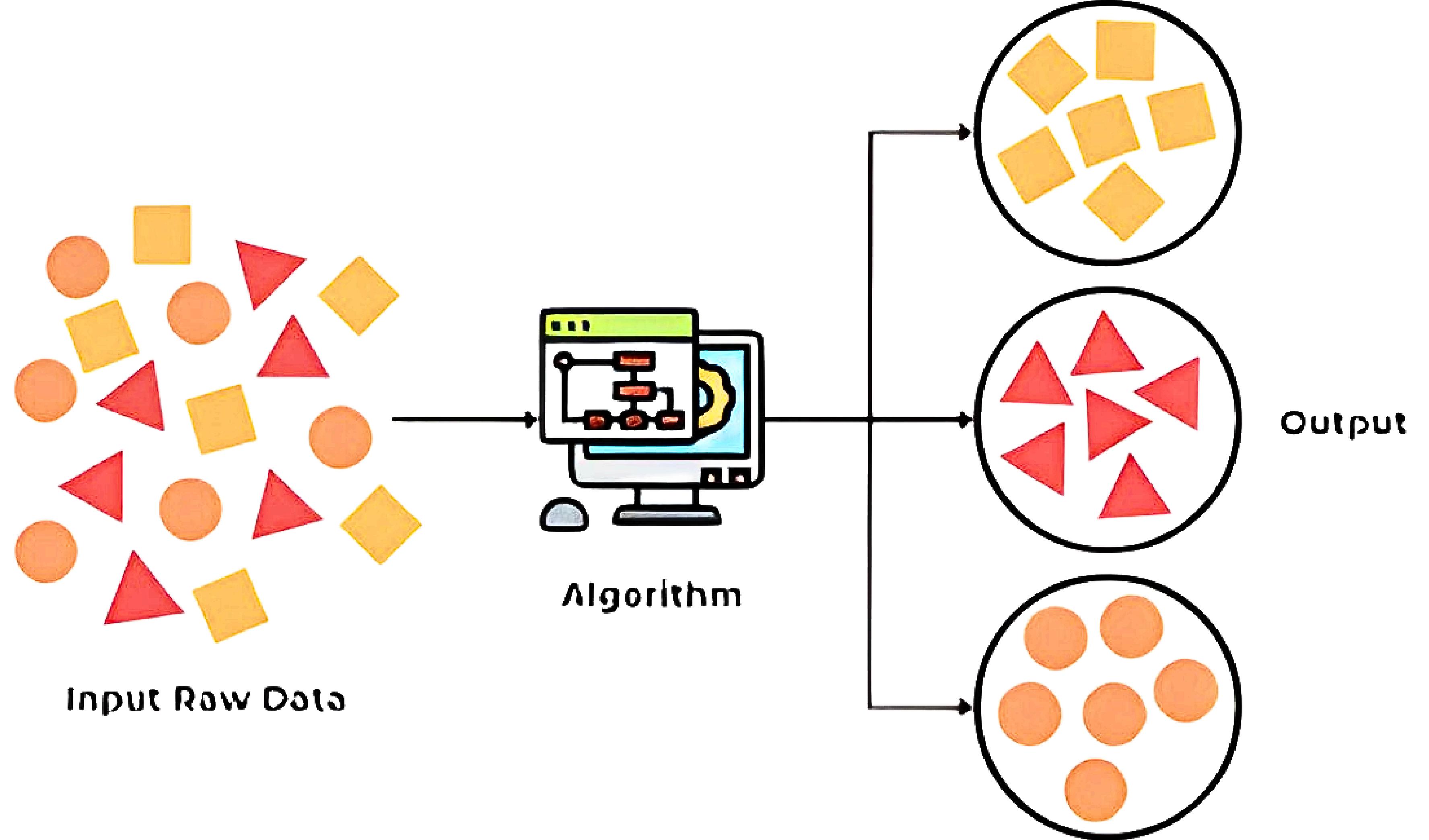
04. Cas du K-Means

05. Application



APPRENTISSAGE NON SUPERVISÉ





L'apprentissage non-supervisé

est une catégorie générale d'algorithmes où le modèle est entraîné sur des données non étiquetées pour découvrir des structures, des modèles ou des relations intrinsèques. Le clustering est l'une de ces techniques d'apprentissage non supervisé.

Faire de l'apprentissage non supervisé, c'est rechercher des relations, des similarités sur un jeu de données sans avoir d'étiquettes !

J'ai des plumes + j'ai des ailes = **espèce X** sans savoir que je suis un oiseau.

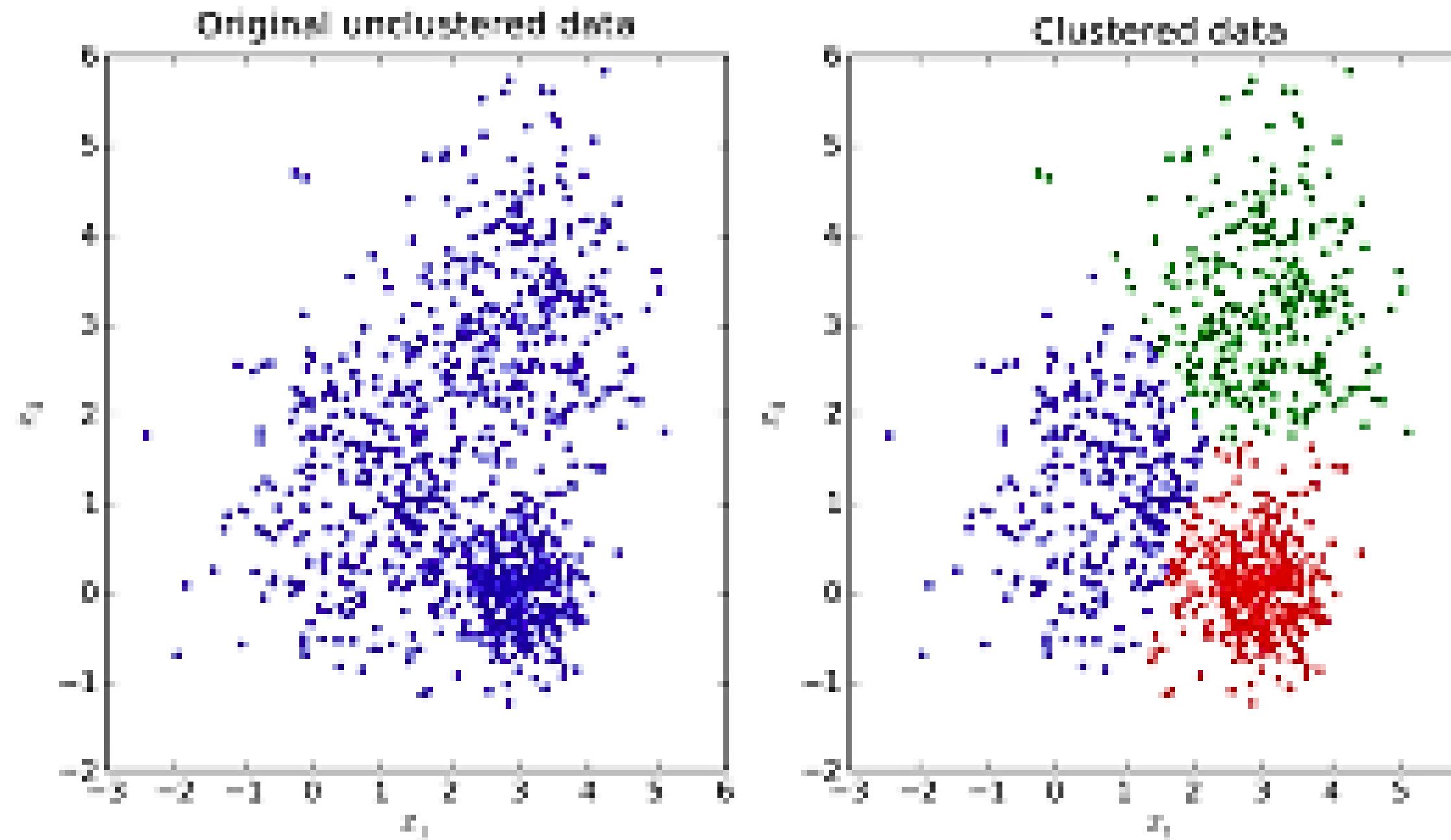
J'ai des poils + j'ai 4 pattes = **espèce Y** sans savoir que c'est un mammifère.

Si les données avaient été étiquetées, on aurait eu "**oiseau**" et "**mammifère**" directement.

CAS D'USAGES



Clustering

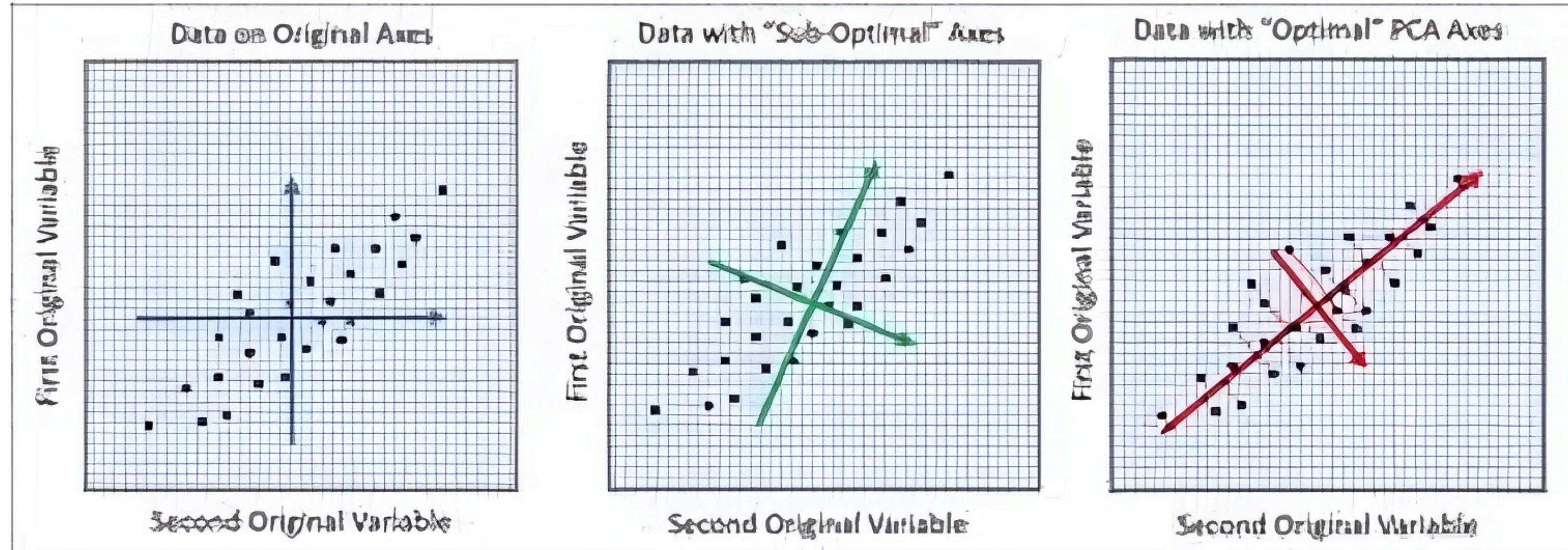


K-Means

Clustering
Hierachique

DBSCAN

Réduction de Dimensionnalité



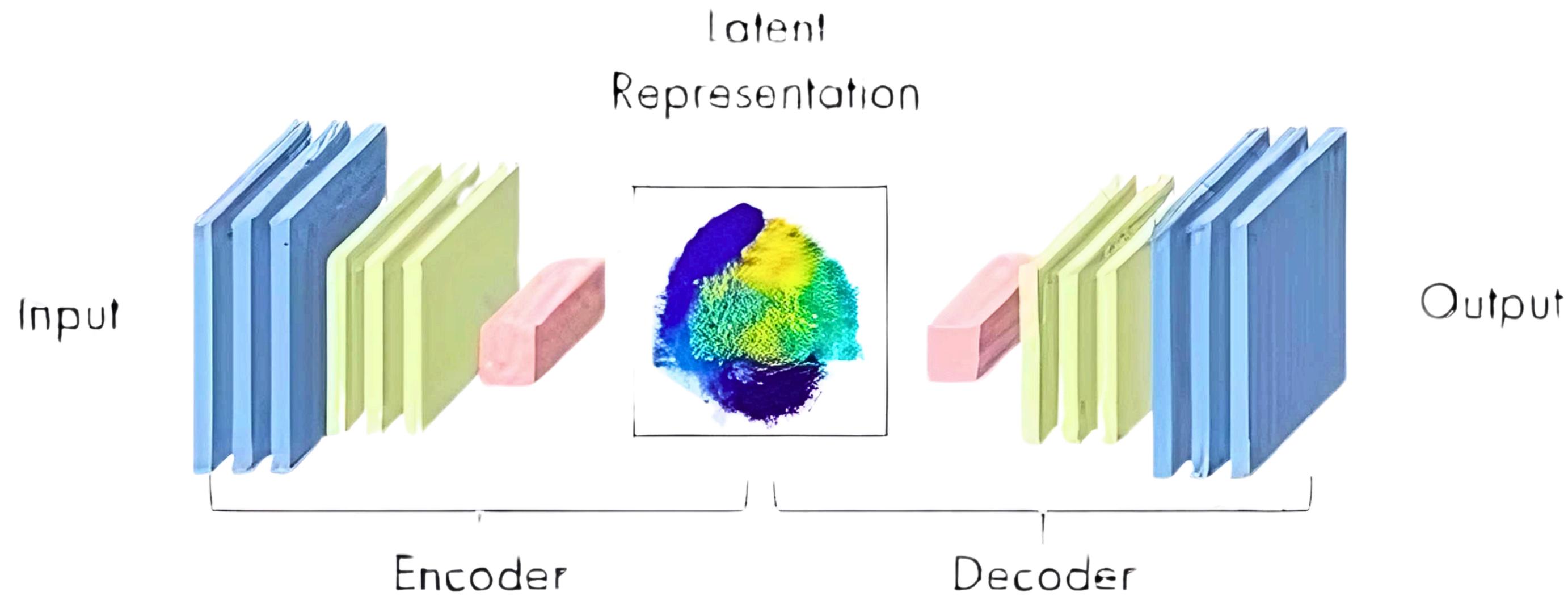
PCA
(Principal Component Analysis)

t-SNE
(t-Distributed Stochastic Neighbor Embedding)

Autoencoders

Réseaux de Neurones Non Supervisés

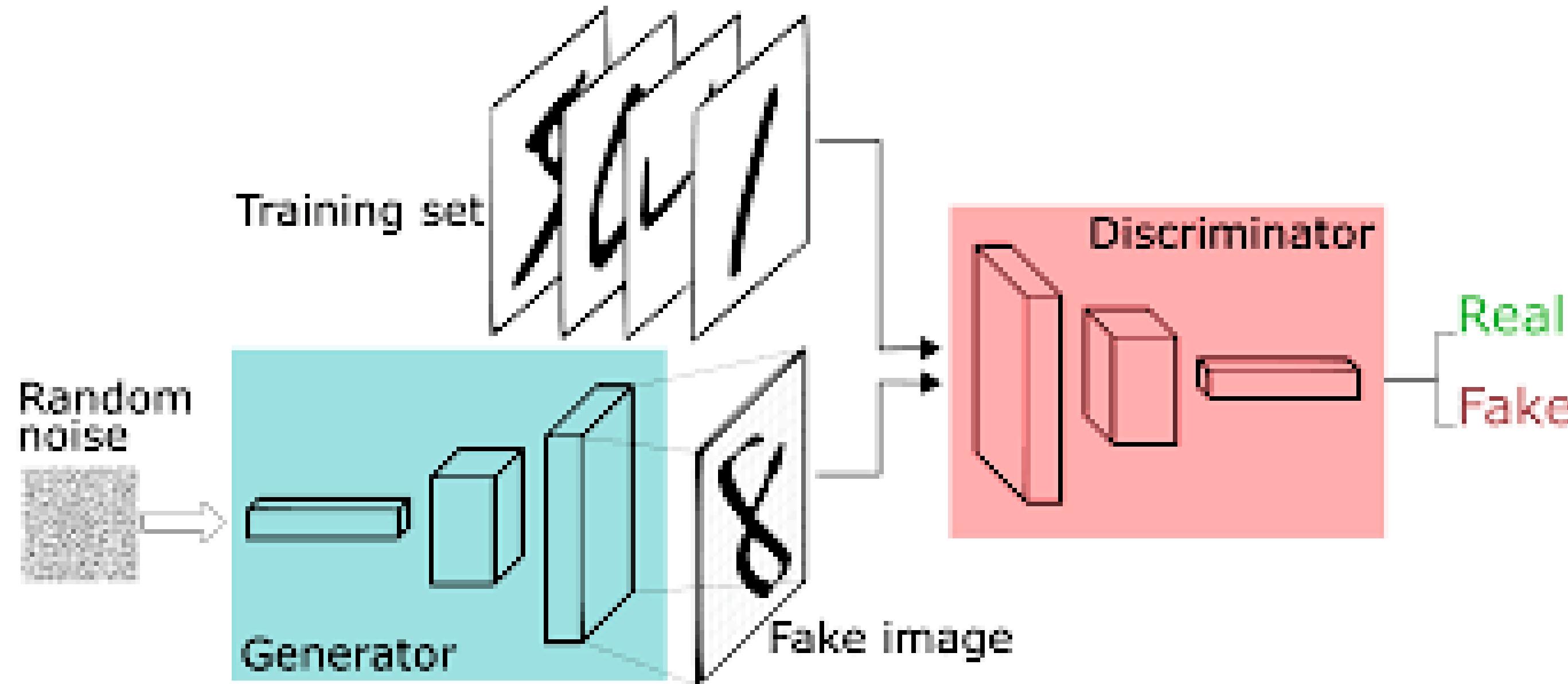
AutoEncoders



Réseaux de neurones entraînés à encoder les données d'entrée en une représentation plus compacte puis à les décoder, utilisés pour la réduction de dimensionnalité, la génération de données, et la détection d'anomalies.

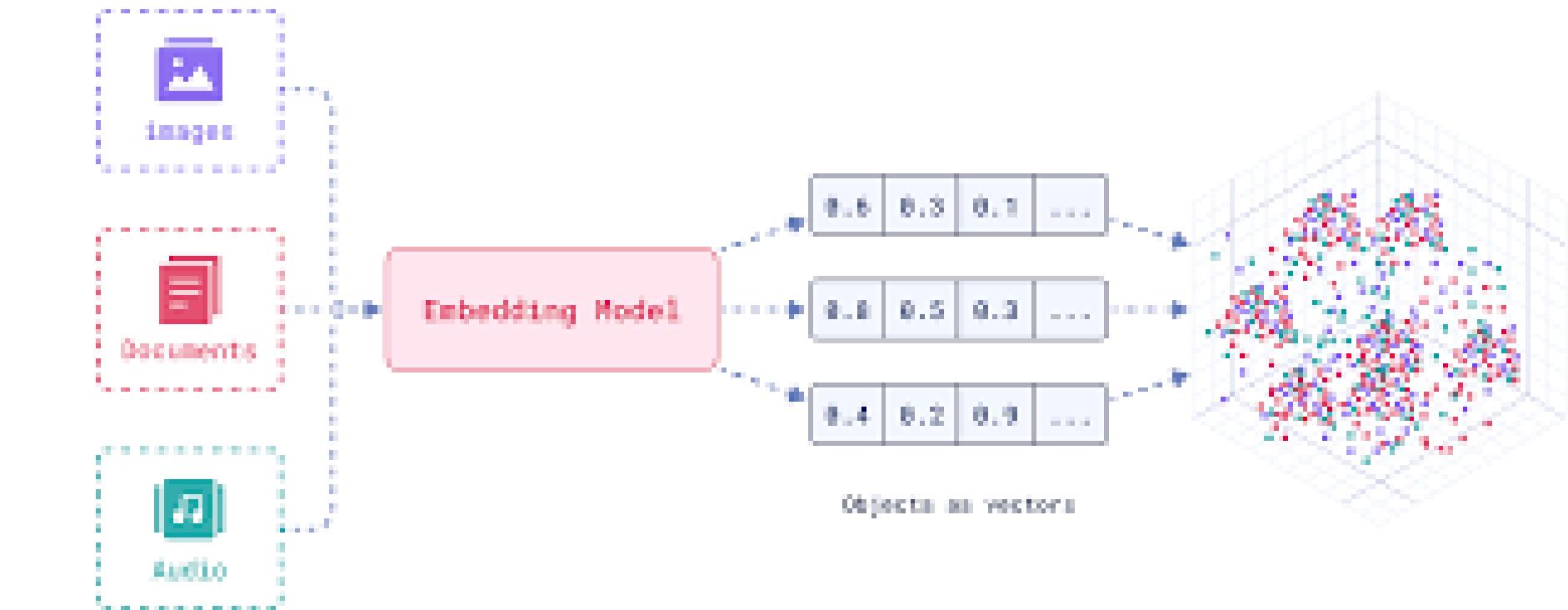
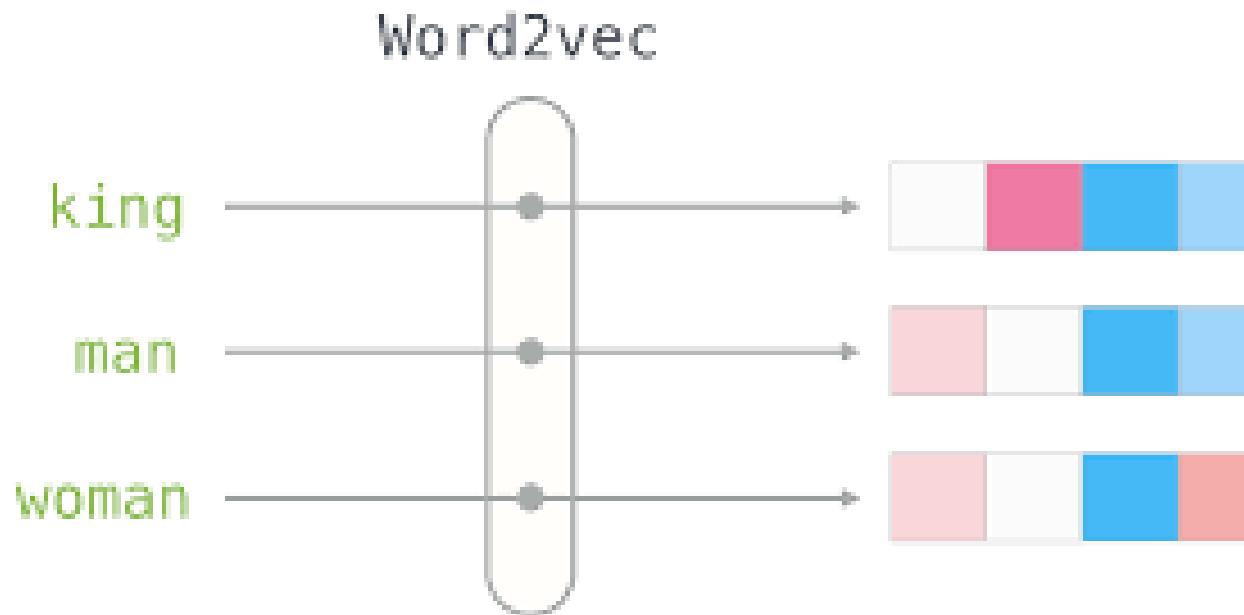
Réseaux de Neurones Non Supervisés

Generative Adversarial Networks (GANs)



Modèle génératif non supervisé composé de deux réseaux (un générateur et un discriminateur) qui s'affrontent pour créer des données réalistes.

Apprentissage par Représentation



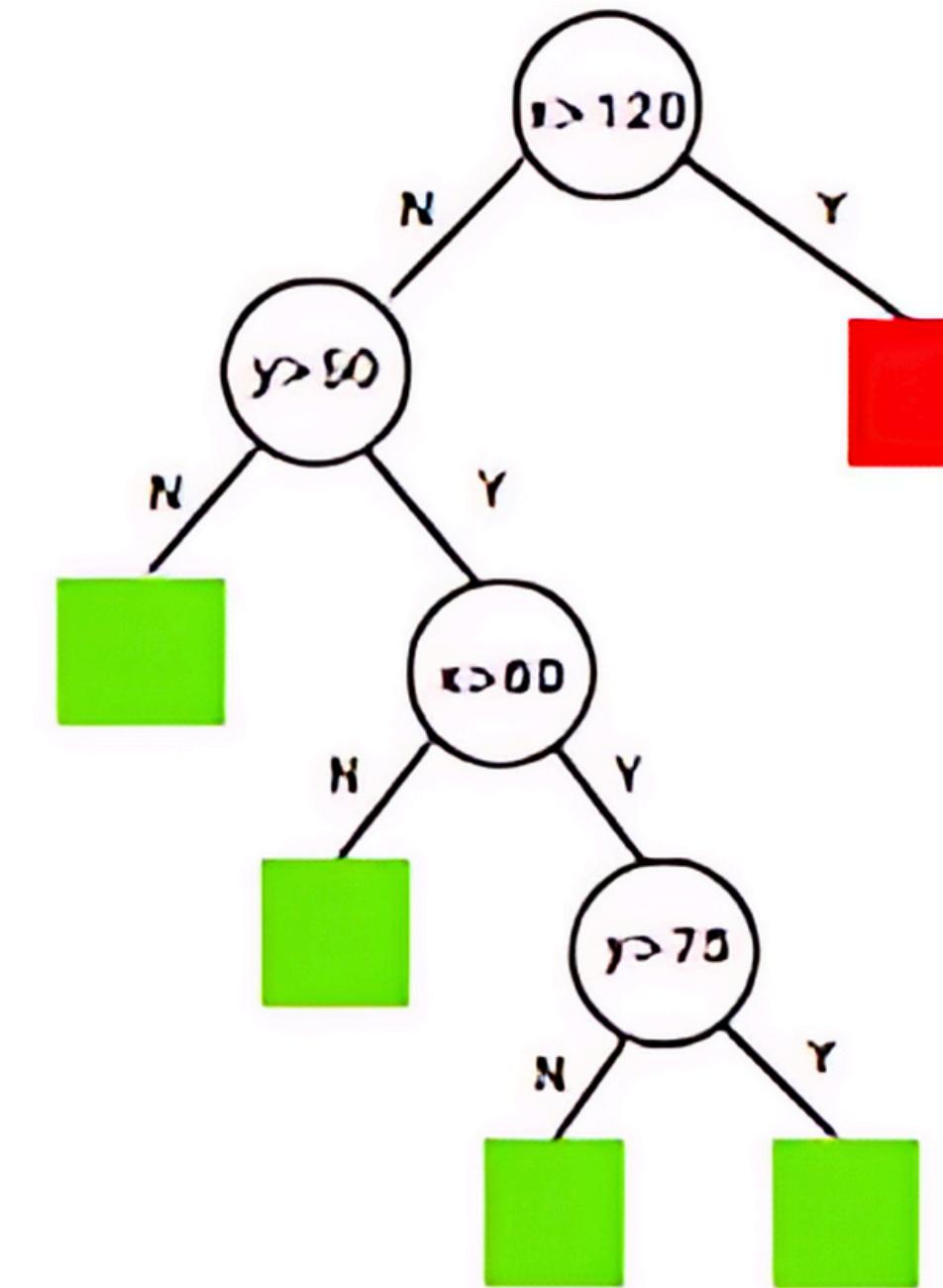
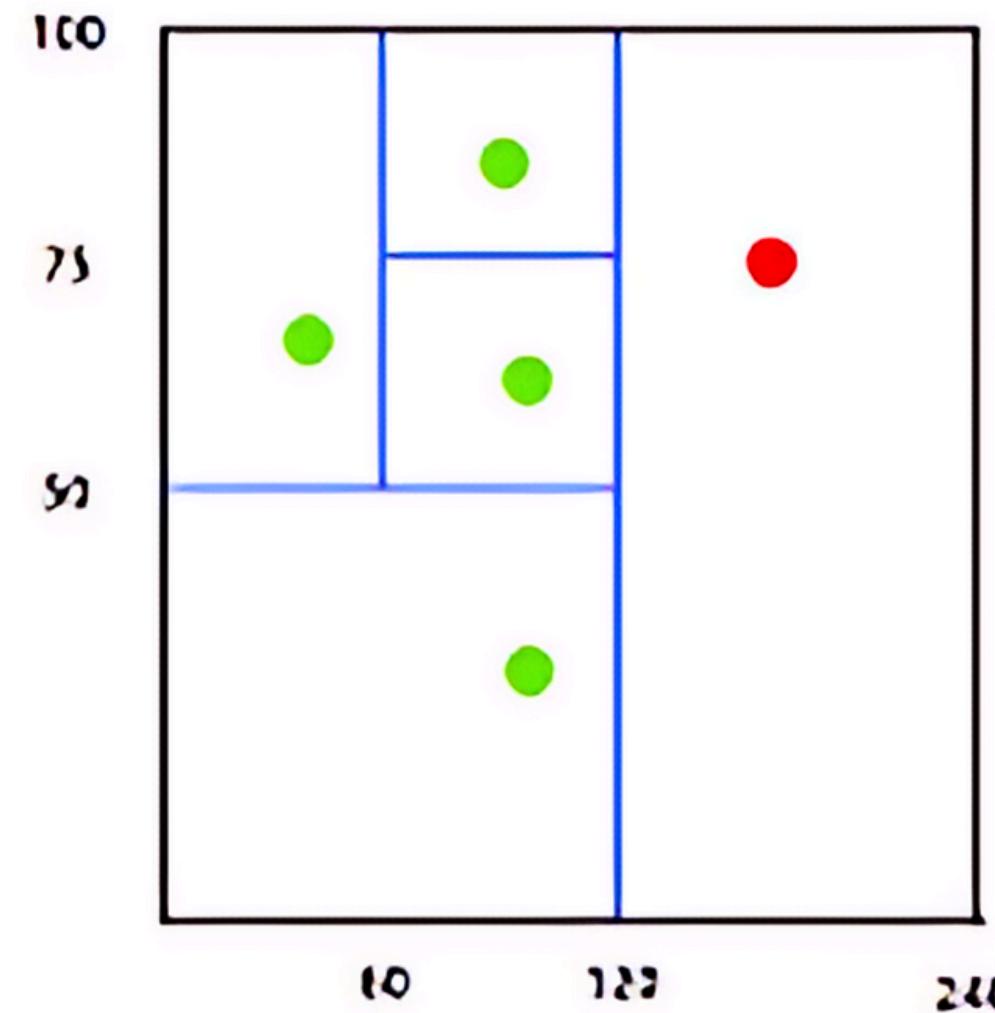
Word2Vec

Embeddings

Modèle pour l'apprentissage de représentations vectorielles des mots à partir de corpus de texte non étiqueté.

Détection d'anomalies

Isolation Forest



Algorithme d'ensemble pour la détection d'anomalies basé sur la construction d'arbres d'isolement.

Classical Machine Learning



Obj: Predictions & Predictive Models

Pattern/ Structure Recognition

CLUSTERING



Le Clustering c'est quoi?

Papa ,je pensais qu'un homme ne pouvait être que noir!
mais j'ai rencontré des hommes blanc et même rouge
l'humanité est caractérisé par une diversité d'espèces.

Aujourd'hui que tu es capable de les distinguer alors garde
pour l'instant qu'on associera le noir à l'Afrique,le blanc à
l'occident et le rouge à l'ancienne Amérique.

C'est ça le **Clustering**. Au début de l'étude, on ignore ce que chaque humain (**élément**) représente; on connaît juste ses **caractéristiques** et on s'en sert pour les distinguer. Ensuite, en créant les groupes, appelés **clusters**, on consulte un expert métier (papa) pour nous aider à associer chaque **groupe** à une **étiquette**.

À titre de comparaison, lors de la classification en apprentissage supervisé, on connaît dès le début le nom de chaque classe.

Avant Clustering

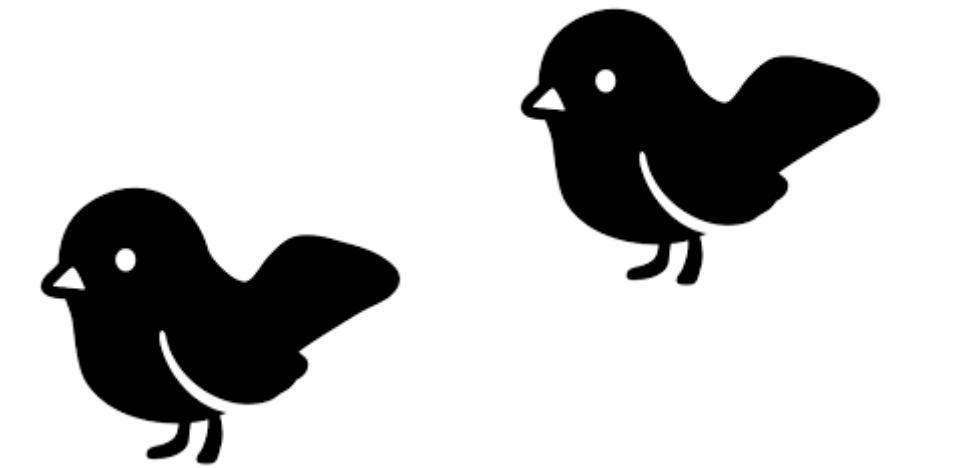


Population

Après Clustering

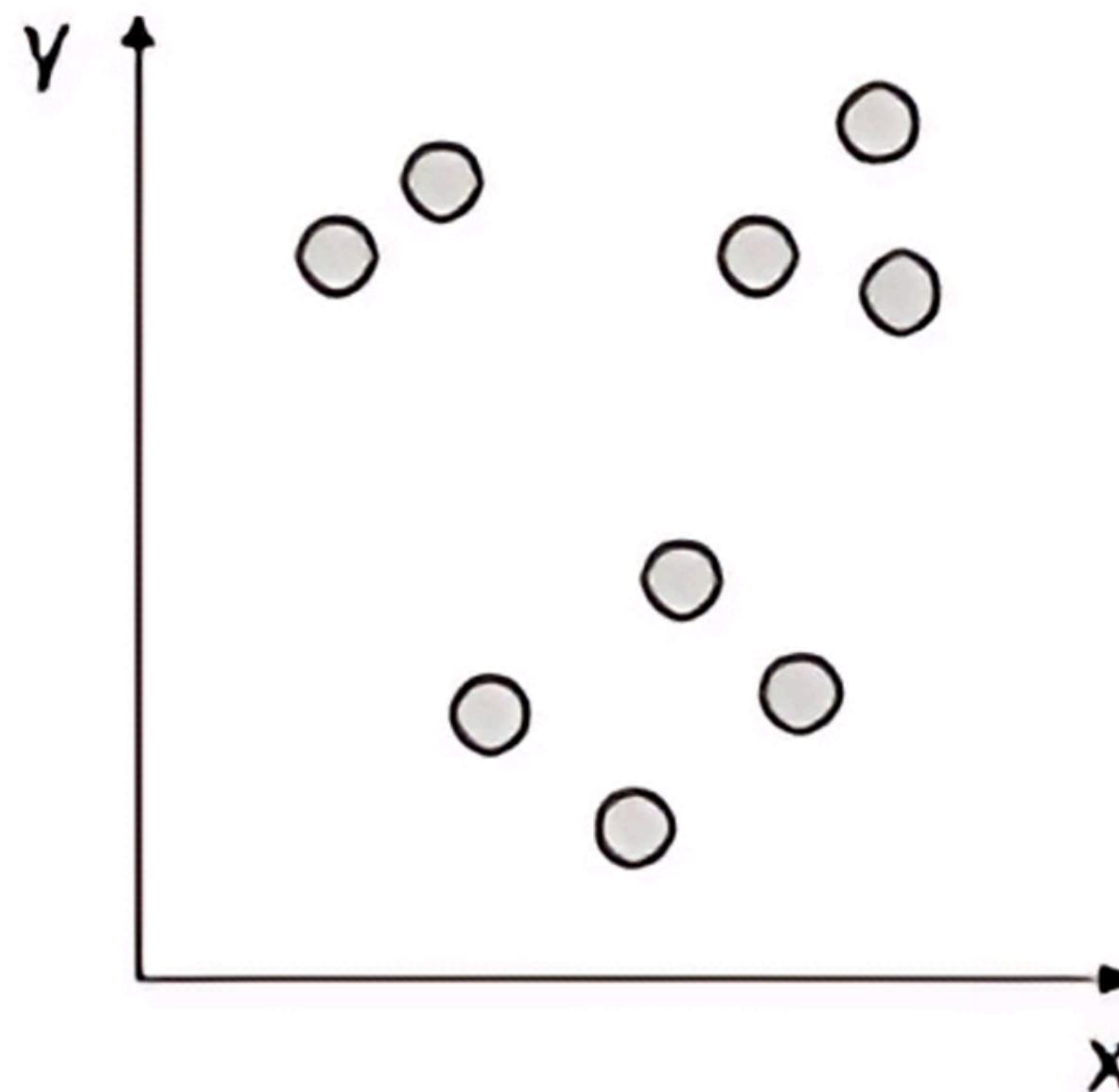


Espèce Y



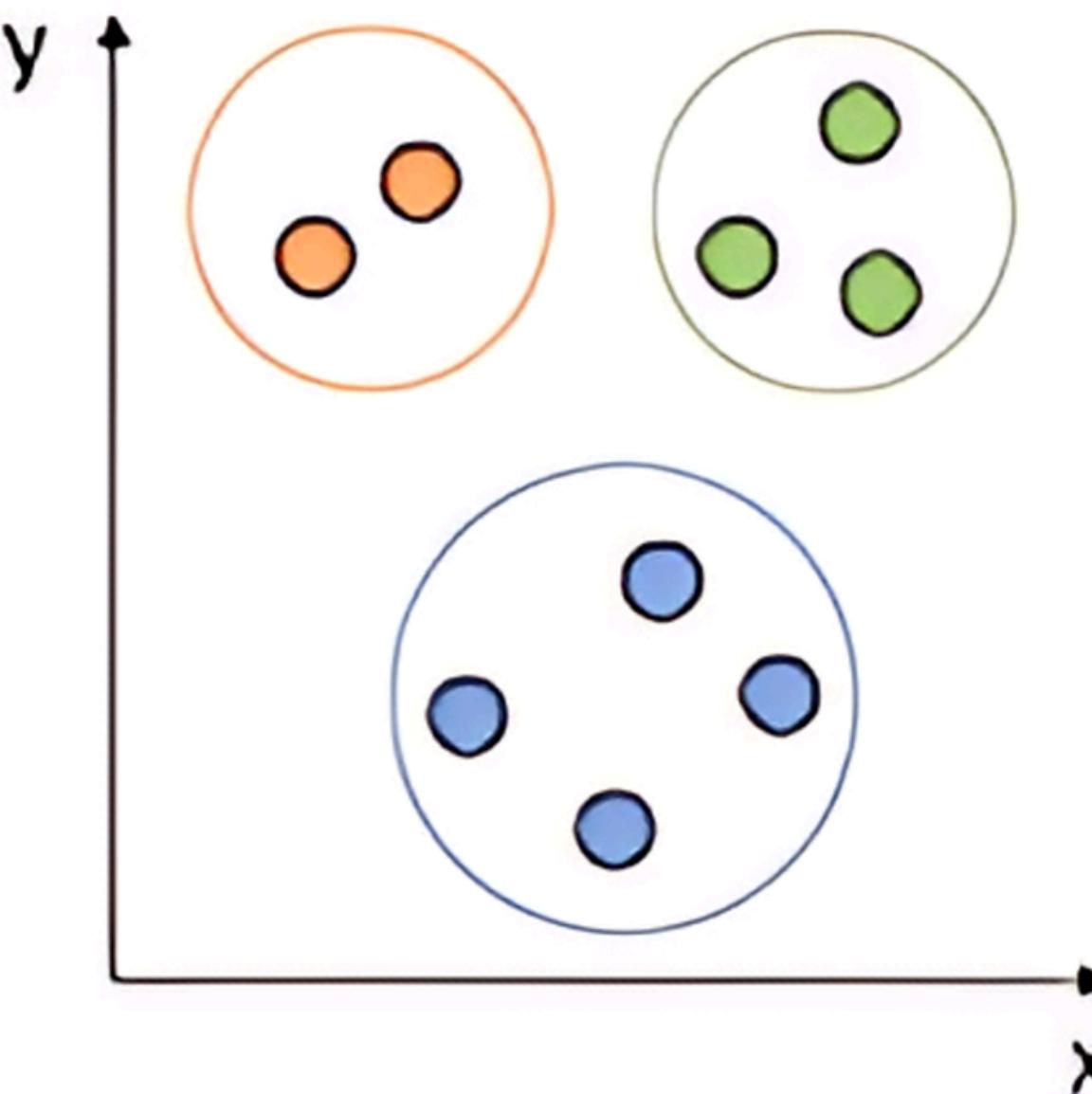
Espèce X

Original Data



Clustering

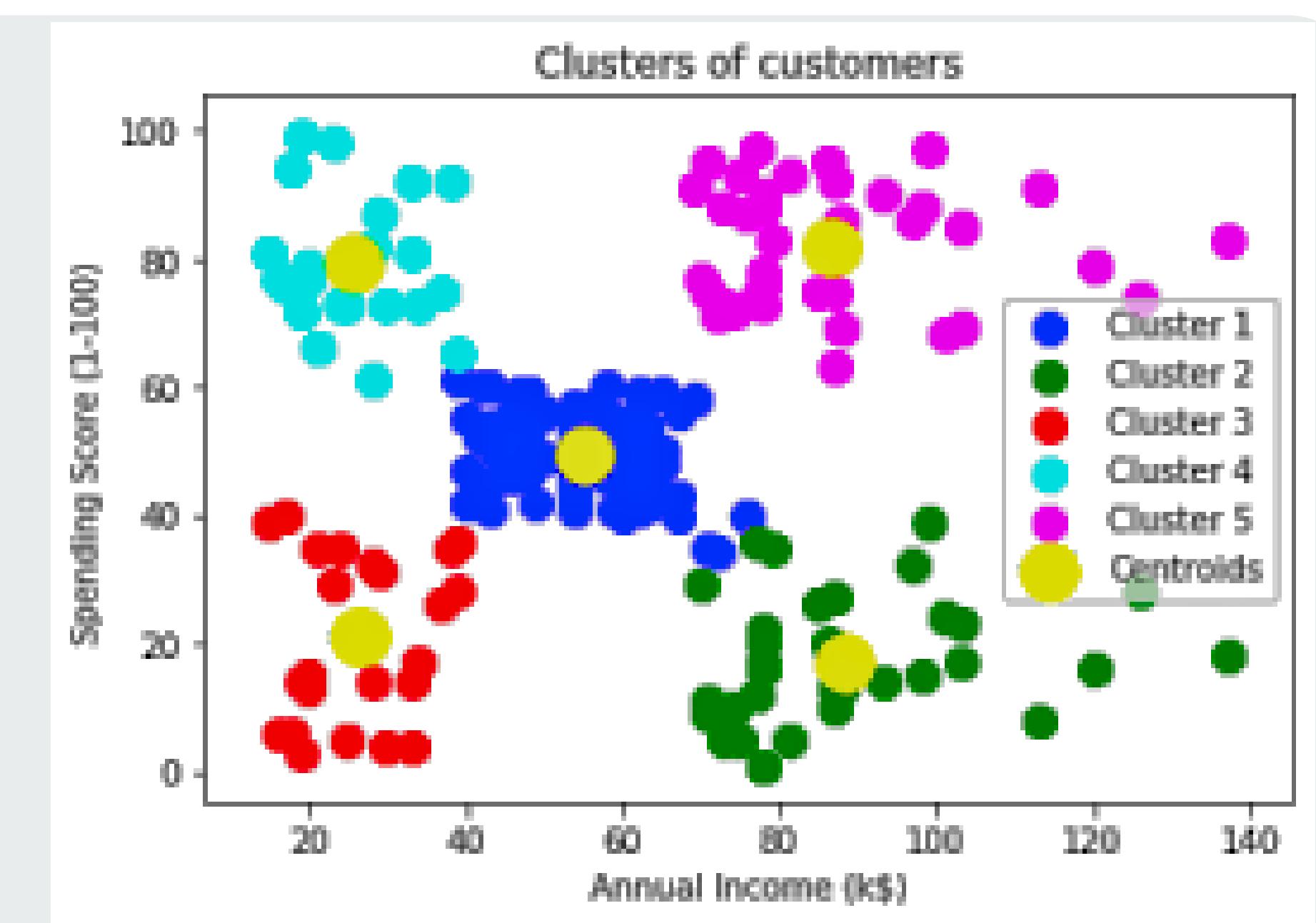
Clustered Data



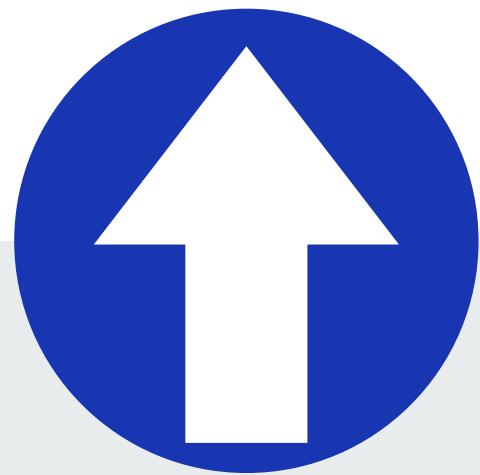
ALGORITHMES DE CLUSTERING

K-Means

Algorithme de clustering itératif où les données sont divisées en k clusters, chaque cluster étant représenté par la moyenne (centroïde) de ses points.



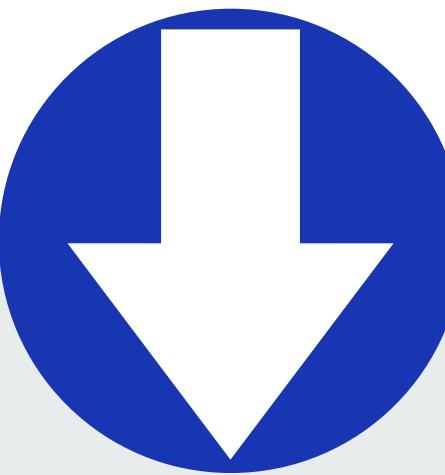
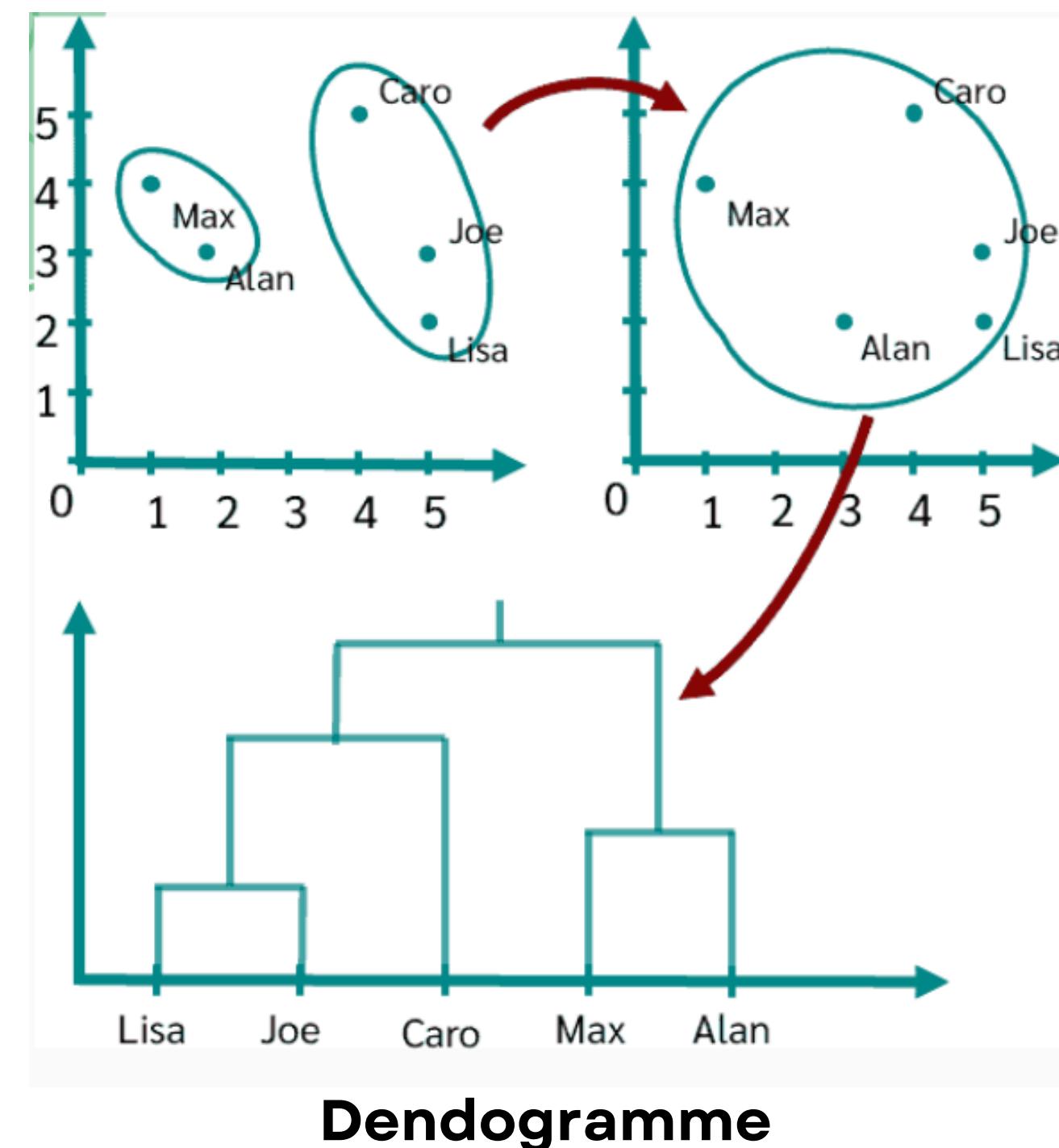
ALGORITHMES DE CLUSTERING



Clustering Agglomératif

Technique ascendante où chaque point commence dans son propre cluster, et des clusters sont ensuite fusionnés par paires jusqu'à ce qu'il ne reste qu'un seul cluster.

Clustering Hiérarchique



Clustering Divisif

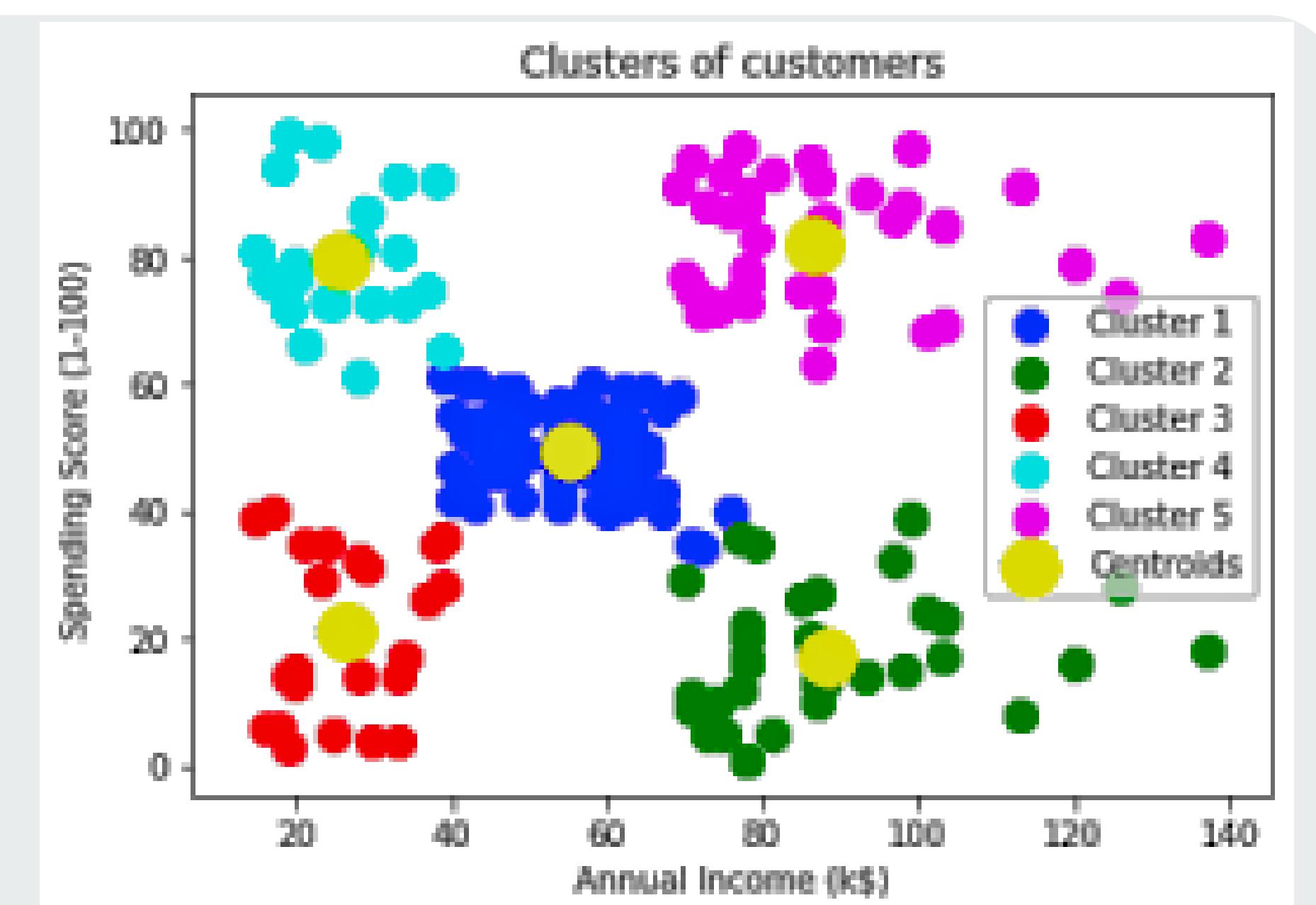
Technique descendante où l'on commence par un seul cluster regroupant toutes les données, puis on le divise successivement en sous-clusters.

ALGORITHMES DE CLUSTERING

DBSCAN

(Density-Based Spatial Clustering of Applications with Noise)

Identifie des régions de densité élevée et les isole en tant que clusters tout en marquant les points peu denses comme bruit.



Plus profondément, Comment
fonctionne le K-Means



The background features two abstract white line art patterns on a dark blue background. On the left, there is a network graph composed of numerous small, irregular polygons and connecting lines, with several small black dots representing vertices. On the right, there is a hierarchical binary tree structure with multiple levels of branches and leaves.

CLUSTERING K-MEANS

Le k-means, algorithme majeur de clustering, vise à diviser les données en k clusters homogènes distincts, où k est un nombre prédéfini.

Alors comment ça marche?

L'objectif est de minimiser la variance intra-cluster tout en maximisant la variance inter-cluster.

La variance intra-cluster est une mesure de la dispersion des données à l'intérieur de chaque cluster. Elle représente la somme des carrés des distances entre chaque point de données d'un cluster et le centre de ce cluster.

La variance inter-cluster elle est définie de manière presque identique à la variance intra-cluster. La différence est que vous ne calculez pas la variance entre tous les échantillons à l'intérieur d'un seul cluster, mais vous prenez le centroïde de chaque cluster (typiquement la moyenne de tous les échantillons à l'intérieur d'un cluster) et calculez la variance entre tous les centroïdes¹. En d'autres termes, elle mesure à quel point les centres des différents clusters sont dispersés.

Etapes du K-Means

1

Initialisation

Certains algorithmes nécessitent une initialisation des centres des clusters(. Choisir de manière aléatoire ou utiliser une méthode telle que K-means++ pour déterminer les centres initiaux des clusters.)

2

Assignation

Assignez chaque point de données au cluster dont le centre est le plus proche. Ici le centre est désigné par CENTROIDES.
Cela se fait généralement en utilisant une mesure de distance, souvent la distance euclidienne.

3

Mise à jour des centres

Recalculez les centres des clusters en prenant la moyenne des points de données assignés à chaque cluster.

4

Répétez

Répétez les étapes 2 et 3 jusqu'à ce que les centres des clusters convergent ou que le nombre d'itérations spécifié soit atteint.

5

Fin

Les centres des clusters et
l'assignation finale des
points de données
définissent les clusters.

APPLICATION

Le Notebook 

Le plus gros problème?

Certainement une très grande sensibilité aux données aberrantes et à l'initialisation des centroides.C'est pour cette raison que que d'autres algorithmes existent pour réussir ce clustering.Entre autres,il y a le DBSCA,HDBSCAN,OPTICS etc., qui sont beaucoup plus robuste qu'un k-means ordinaire.